Story Link Detection Based on Event Model with Uneven SVM

Xiaoyan Zhang, Ting Wang, and Huowang Chen

Department of Computer Science and Technology, School of Computer,
National University of Defense Technology
No.137, Yanwachi Street, Changsha, Hunan 410073, P.R. China
{zhangxiaoyan,tingwang,hwchen}@nudt.edu.cn

Abstract. Topic Detection and Tracking refers to automatic techniques for locating topically related materials in streams of data. As a core of it, story link detection is to determine whether two stories are about the same topic. Up to now, many representation models have been used in story link detection. But few of them are specific to stories. This paper proposes an event model based on the characters of stories. This model is used for story link detection and evaluated on the TDT4 Chinese corpus. The experimental results indicate that the system using the event model achieves a better performance than that using the baseline model. Furthermore, it shows a larger improvement to the former, especially when using uneven SVM as the multi-similarity integration strategy.

Keywords: story link detection, event model, uneven SVM.

1 Introduction

Topic Detection and Tracking (TDT) [1] refers to a variety of automatic techniques for discovering and threading together topically related materials in streams of data such as newswire or broadcast news. As a core of TDT, story link detection is the task of determining whether two stories are about the same topic. TDT defines “topic” to mean a specific event or activity plus directly related events or activities. An event is "something that happens at some specific time and place".

A number of works have been developed on story link detection, which can be classified into two categories: vector-based methods and probabilistic-based methods.

As the vector-based approaches are widely used in IR and Text Classification research, cosine similarity between documents vectors with *tf***idf* term weighting[2,3] has also been one of the best methods for link detection. We have also examined a number of similarity measures in the link detection task, including weighted sum, language modeling, Kullback-Leibler divergence, Hellinger and Tanimoto, and find that cosine similarity produces the most outstanding results. Furthermore, [4] also confirms this conclusion in its story link detection research. Probabilistic-based methods have been proved to be very effective in several IR applications. One of their attractive features is that it is firmly rooted in the theory of probability, thereby allows the researcher to explore more sophisticated models guided by the theoretical framework. [5,6] both apply probability-based models (language model or relevance model) to

H. Li et al. (Eds.): AIRS 2008, LNCS 4993, pp. 436–441, 2008.

story link detection, and the experimental results indicate that the performances are comparable with those using traditional vector space models, if not better. The story models are all vector-based in this paper. We have concluded in our previous research that the multi-vector model is superior to the single-vector model for news stories. So a multi-vector model is used as the baseline model. However, we know that a story usually describes an event mainly constructed with time, location, person, organization, etc. So a new event model is proposed to represent the news story. The experimental results show that the event model is more proper for stories in TDT.

The paper is organized as follows: Section 2 simply describes the procedure for story link detection; Section 3 describes the baseline multi-vector model and the event model, which share preprocessing, feature weighting, similarity computation and multi-similarity integration except model construction. The experimental results and analysis are given in Section 4; finally, Section 5 concludes the whole paper.

2 Problem Definition

In story link detection, a system is given by a sequence of time-ordered news source files $S=\langle S_1, S_2, ..., S_n\rangle$, each S_i $(i=1, 2, \ldots, n)$ includes a set of stories and a sequence of time-ordered story pairs $P=\langle P_1, P_2, ..., P_m\rangle$, where $P_i= (s_{i1}, s_{i2})$, $s_{i1}\in S_j$, $s_{i2}\in S_k$, $1\le i\le m$, $1\le j\le k\le n$. The link system is required to make decisions on each story pair to judge if two stories in it are topically linked. The procedure of processing a story pair is as follows. For current story pair $P_i= (s_{i1}, s_{i2})$:

1. Get background corpus B_i of P_i. Normally, according to the supposed application situation, the system is allowed to look ahead N (usually 10) source files when deciding whether the current pair is linked. So,

$$B_i = \{S_1, S_2, ..., S_l\}, \text{ where } l = \begin{cases} k+10, & s_{i2} \in S_k \ \textit{and} \ (k+10) \le n \\ n, & s_{i2} \in S_k \ \textit{and} \ (k+10) > n \end{cases}.$$

2. Produce the representation models (M_{i1}, M_{i2}) for two stories in P_i. $M= \{(f_s, w_s)| s\ge 1\}$, where f_s is a feature extracted from the story and w_s is the weight of the feature in the story. They are computed with some parameters counted from current story or the background.
3. Choose a similarity function F and compute the similarity between two models. If $F(M_{i1}, M_{i2})$ is larger than a predefined threshold, then two stories are decided to be topically linked.

3 Baseline Model and Event Model

First of all, a preprocessing is provided for all the stories. For each story we tokenize the text, tag the tokens, remove stop words, and get a candidate set of features for its vector-based model. In this paper, a token plus its tag is taken as a feature. If two tokens with same spelling are tagged with different tags, they will be taken as

different features. The segmenter and tagger are completed by ICTCLAS[1]. The stop word list contains 507 words.

3.1 Model Construction

In the baseline model, we divide the feature set into disjoint subsets according to the tags of tokens in the set. One vector represents a subset. After that ten vectors are picked out to represent the story. The ten corresponding tags are person name, location name, organization name, number, time, noun (including noun and proper noun), verb (including verb, associate verb and nounness verb), adverb, adjective (including adjective, associate adjective and nounness adjective), idiom (including idiom and phraseology). We think that those tokens, which are tagged with any of these ten tags, have comparably more information. So the baseline multi-vector model is actually a ten-vector model.

However, we know that the research objects in TDT are news stories, not usual documents. Stories have their own features besides the usual characters. For example, almost each story describes an event. The time, location, person, and so on, compose the framework of an event. And also the first few sentences often summarize the event and the rest sentences explain what the event is about in detail. Story link detection is to decide whether two stories are topically linked. The topic here is the event described in a story and other event directly related to it. Topic here is event based. Therefore the representation model for stories should reflect the events described in them from both content and structure. The event frame maybe a more proper and natural partition standard for a multi-vector model to represent a story. In this paper we build a new multi-vector model for a story according to the event framework. We call it event model. The later experimental results also verify that just splitting the feature set according to the tag is not the best choice. The event model gets much improvement to the performance of story link detection and has also a larger potentiality for improvement than the baseline model.

The main difference between the event model and the baseline model is the partition standard of the feature set. The baseline model is built according to the tags while the event model is built according to the event framework. The primary component elements are time, number, person, location, organization, abstract and content description in our event framework. The first five will be the same as those in the baseline model. The features in the abstract vector and the content description vector are tokens with their tags individually occurring in the first two sentences and the rest sentences in a story, while they do not belong to the former five kinds of tokens. It is because we find that the first two sentences in a story usually summarize the story and the rest sentences explain what happened in detail. According to this splitting standard, the feature set will be split into seven subsets.

3.2 Other Common System Components

Firstly, the weights of the features in above two models are all based on the *tf*idf* form. Furthermore, it is dynamic, which lies in the dynamic computation of some

[1] http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/

parameters in the *tf*idf* form. The story collection used for statistics is incremental, since more story pairs are processed, more source files could be seen, and the story background corpus is bigger. Whenever the size of the story background has changed, the values of some certain parameters will update accordingly. We call this incremental *tf*idf* weighting.

Secondly, another important issue is to determine the right function to measure similarity between two corresponding sub-vectors. The critical property of the similarity function is its ability to separate vectors describing the same information from vectors on different information. We consider the classical cosine function in this paper. This measure is simply an inner product of two vectors, where each vector is normalized to unit length. It represents cosine of the angle between two vectors.

The last important step for the baseline model and the event model is to integrate multiple similarities into a single value to decide whether two stories are topically linked. We do this with a machine learning classifier SVM in this paper. It is because SVM has a good generalization property and has been shown to be a competitive classifier for a variety of other tasks [7]. Firstly, SVM is trained on a set of labeled vector where the features are similarities between two corresponding subvectors and the topically link label in each vector. The generated model is then used to automatically decide whether two stories in a new pair are linked. We use the SVMlight[2] software in this paper. A radial basis function is used in all the reported experiments. In addition to making a decision for whether two stories are linked, SVM also produces a value as the measure of confidence, which is served as input to the evaluation software.

We also notice that the numbers of positive and negative examples in the training data are very different. But the usual SVM treats positive and negative data equally, which may result in poor performance when applying to very unbalanced detection problems. [8] presents a method to solve this problem, where the cost factor for positive examples is distinguished from the cost factor for negative examples to adjust the cost of false positive vs. false negative. This approach is implemented by the SVMlight, in which an optional parameter j ($=C_+/C_-$) is provided to control different weightings of training errors on positive examples to errors on negative examples. Therefore, we also integrate a set of similarities with uneven SVMlight in this paper.

4 Experiment

We use the Chinese subset of TDT4 corpus in this paper. There are totally 12334 story pairs extracted for our experiments. The answers for these pairs are based on 28 topics in TDT 2003 evaluation. The first 9000 pairs are used for training. The rest 3334 pairs are used for testing. The goal of link detection is to minimize the detection cost (C_{det}) due to errors caused by the system, which is a function of the miss probability (P_{miss}) and the false alarm probability (P_{fa}). The cost for each topic is equally weighted and normalized so that the normalized value ((C_{det})$_{norm}$) can be no less than one. The detailed explanation can refer to [1].

[2] http://svmlight.joachims.org/

4.1 Experimental Results

To verify the effectiveness of the event model and the uneven SVM strategy, we have designed four story link detection systems: System1 uses the baseline model with even SVMlight, System2 uses the baseline model with uneven SVMlight, System3 uses the event model with even SVMlight, and System4 uses the event model with uneven SVMlight. All the systems are implemented according to the procedure introduced in the section of problem definition. When the generated SVM classify model is used to automatically decide whether two stories in a new pair are linked, the default optimum threshold is zero. The following table shows the topic-weighted experimental results of these systems.

Table 1. Topic-Weighted Experimental Results of Four Systems

	P_{miss}	P_{fa}	C_{det}	$(C_{det})_{norm}$
System1	0.0429	0.0048	0.0013	0.0664
System2	0.0149	0.0090	0.0012	0.0588
System3	0.0267	0.0052	0.0010	0.0524
System4	0.0193	0.0060	0.0010	0.0487

From the table we can see that story link detection using the event model has a lower detection cost than that using the baseline model whenever using even or uneven SVM. This may be because the event model does no discriminate those features which are not named entities (person, location, organization, time, number) and just splits them into abstract and content description vectors. On one hand it avoids the cost loss caused by exact matching when comparing two sub vectors. On the other hand it reflects the character that the first two sentences in a story are usually summary of what happened and the rest are explanation of what happened.

Relative to even SVM, uneven SVM has made a notable decrement in P_{miss} while a little increment in P_{fa}. This is because uneven SVM is completed under the principle of decreasing the loss errors at the cost of little increment in false alarm errors. We have verified through experiments that the event model is always superior to the baseline model no matter what the optional parameter j is. So we think that the event model is more appropriate to represent news stories in TDT.

Although the event model has got a comparable lower detection cost, we should also notice that the information of relation between features in event model has not yet been abstracted and used. Only using information of tokens and tags may be insufficient. If we could get the relation between the event in a story and the seminal event of its corresponding topic, we should be able to make the right decision at a larger confidence. We will try to exploit relation information and use them properly in our future work.

5 Conclusion

Story link detection is a key technology in TDT research. Though many models have been used, few of them are specific to stories in TDT. After analyzing the characters

of stories in TDT corpora, this paper proposes a new event model. It represents the events in a story according to the event framework. The experimental results indicate that story link detection using this model can achieve a better performance than that using the baseline model, especially when using the uneven SVM integration strategy. So we think the event model is more proper for TDT stories. However, we realize that the event model is only used in a simple way in this paper, just looking it as a multi-vector model. How to mine and utilize the relation information between features in the event model will be our future work.

Acknowledgement

This research is supported by the National Natural Science Foundation of China (60403050), Program for New Century Excellent Talents in University (NCET-06-0926) and the National Grand Fundamental Research Program of China (2005CB321802).

References

[1] Allan, J.: Introduction to topic detection and tracking. In: Allan, J. (ed.) Topic Detection and Tracking - Event-based Information Organization, pp. 1–16. Kluwer Academic Publisher, Dordrecht (2002)
[2] Connell, M., Feng, A., Kumaran, G., Raghavan, H., Shah, C., Allan, J.: Umass at tdt 2004. In: TDT 2004 Workshop (2004)
[3] Chen, F., Farahat, A., Brants, T.: Multiple similarity measures and source-pair information in story link detection. In: HLT-NAACL, pp. 313–320 (2004)
[4] Allan, J., Lavrenko, V., Malin, D., Swan, R.: Detections, bounds, and timelines: Umass and tdt–3. In: Proceedings of Topic Detection and Tracking (TDT–3), pp. 167–174 (2000)
[5] Nallapati, R.: Semantic language models for topic detection and tracking. In: HLT-NAACL (2003)
[6] Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., Thomas, S.: Relevance models for topic detection and tracking. In: Proceedings of Human Language Technology Conference (HLT), pp. 104–110 (2002)
[7] Van Der Walt, C.M., Barnard, E.: Data characteristics that determine classifier performance. In: Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa, pp. 160–165 (2006)
[8] Morik, K., Brockhausen, P., Joachims, T.: Combining statistical learning with a knowledgebased approach - a case study in intensive care monitoring. In: Proceedings of the 16th International Conference on Machine Learning (ICML 1999), pp. 268–277. Morgan Kaufmann, San Francisco, CA (1999)